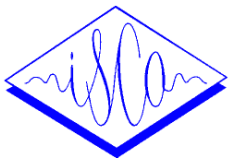

ESCA Workshop on
Audio-Visual Speech Processing
Rhodes, Greece
September 26-27, 2997

ADAPTIVE DETERMINATION OF AUDIO AND VISUAL WEIGHTS FOR AUTOMATIC SPEECH RECOGNITION

Alexandrina Rogozan, Paul Deléglise and Mamoun Alissali
Laboratoire d'Informatique de l'Université du Maine
Université du Maine, 72085 Le Mans Cedex 9, France.
Tel. ++33 (0)2 43 83 38 64, FAX: ++33 (0)2 43 83 38 68, E-mail: afoucaul@lium.univ-lemans.fr

ABSTRACT

This paper deals with adaptive integration of visual information in an automatic speech recognition system. Our method consists of attaching a different weight to each modality involved in the recognition process. These acoustic and visual weights are adjusted dynamically, manly according to the SNR, which is provided to the system as a contextual input.

This method is tested on three different audio-visual CHMMs-based systems. They implement respectively: the direct identification scheme (DI), the separate identification scheme (SI) and the hybrid (DI+SI) one. System performances are compared on the same task: speaker-dependent continuous spelling of French letters.

Results obtained using audio and visual weights dynamically adapting to the circumstances are better than those obtained with equal weights, over different test condition (clean data and data with artificial noise).

1. INTRODUCTION

Several researchers have already demonstrated, through their models of automatic audio-visual perception, the potential use of visual information (mostly lip shape and movements) to improve accuracy and robustness of speech recognition systems (Adjouani and Benoît, 1996; Meier *et al.*, 1996; Silsbee and Su, 1996).

Our own work in this area (Alissali *et al.*, 1996) was focused on the elaboration of an optimal integration strategy of audio and visual sources in automatic speech recognition. The experiments realised under different test conditions show that a hybrid (DI+SI) identification model is more promising than the separate identification and asynchronous integration (SI) or than the direct integration (DI) and confirm (Robert-Ribes *et al.*, 1996).

However, these results correspond to the empirically-obtained optimal modality weights, i. e. the acoustic and visual weights which give the best performances. These performances underline the importance of the weighting modality factor in the AV recognition process.

Our acoustic-only system achieved recognition accuracy of 90% on clean data, while the visual recogniser performance does not exceed 44 %. These results prove that in non-noisy condition, acoustic source is most reliable and the visual part should be lower in the recognition process. In noisy environment the acoustic system perform poorly and the visual source becomes necessary. Thereby, an optimal integration model has to adapt the relative contribution of each modality according to the SNR.

On the other hand, since auditory and visual confusion of phonemes are manly independent, the recognition errors are modality-dependent. By the way, one has to exploit the auditory and visual confusion for determining optimal modality weights in order to improve recognition performance.

This paper focus on adaptive determination of audio and visual weights manly according to the SNR. This method is tested on audio-visual systems, which implement respectively: the direct identification scheme (DI), the separate identification scheme (SI) and the hybrid (DI+SI) one.

Generally, AV system performances obtained using audio and visual weights dynamically adapting to the circumstances are better than those obtained with equal weights, over different test condition.

2. SYSTEM DESCRIPTION

The baseline acoustic-only system uses phonemic CHMMs, where the acoustic observations are composed of 12 MFCC coefficients, the energy of the analysis window and their first and second derivatives.

Starting from the basic system, we implement the DI model by simply concatenating the acoustic and visual observations in a first AV system. As visual observation, we use parameters representing the internal lip shape (height, width and area), obtained by image processing (Lallouache, 1991), and their first and second derivatives.

In the second AV system developed according to the SI model, the audio and visual observations are proc-

essed in two separate communicating components. Since the visual component is only used to rescore solutions proposed by the acoustic one, the visual information may be discarded.

To avoid this problem, we developed a third AV system which implements a hybrid DI+SI model. We do this by replacing the acoustic component of the second AV system by an DI-based component.

3. ACOUSTIC AND VISUAL WEIGHTS DETERMINING

3.1 Adaptive Weights for Direct Identification Model

In the first AV system, we denote the bimodal observation at time t by $o_t = [a_t, v_t]$ obtained by merging the corresponding acoustic and visual observations. We assume that both observations are statistically independent, which correspond to use a diagonal covariance matrix and by the way reduce the total number of parameters to estimate.

With the previous assumption, the composite observation o_t could be split into two streams. In this manner, the probability to observe o_t at instant t in state i could be written:

$$P(o_t \mid state_t = i) = P_i(a_t) \times P_i^{\gamma_V}(v_t) \qquad (1)$$

where $P_i(a_t)$ and $P_i^{\gamma_V}(v_t)$ represents the probability of the acoustic vector, respectively the probability of visual vector exponentially weighted by γ_V.

At the time of Baum-Welch learning or Viterbi decoding, using the probability estimate (1) allows for change in the visual source contribution, and a bias the estimated likelihood. For γ_V values close to 1, visual source will contribute likewise the acoustic one in the likelihood estimation, while values of γ_V less than 1 attenuate its importance.

The visual weighting factor γ_V could be learn to adapt the training set as in (Silsbee and Su, 1996). Because our training data is not sufficient to adopt this approach, we determinate it as a linear function of the noise level like in (Meier *et al.*, 1996) (see Figure 1).

The phonemic CHMMs are trained on the same utterances of clean and noisy data. During the test, the visual weighting factor γ_V is adjusted according to the noise level, which is provided to the system as a contextual input.

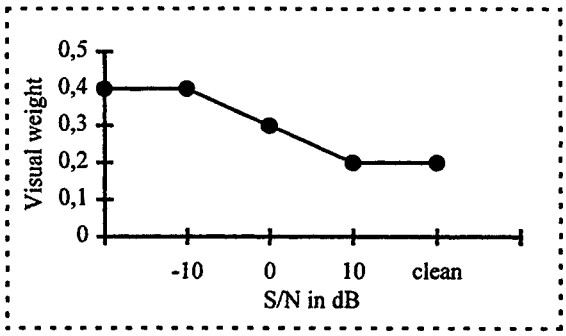

Figure 1: Variation of the visual weighting factor γ_V as a function of noise level in the DI based system

3.2 Adaptive Weights for Separate Identification Model

In the second AV system, the acoustic component furnish N-best acoustic recognition hypotheses. The visual component is used to compute a visual score for each acoustic recognition hypothesis. Since taken into account the correlation between the acoustic and the visual component is relatively complex, we assume the independence between both modalities. Thereby, acoustic and visual scores could be combined in a linear weighting manner:

$$S_{best} = \lambda \times S_v + (1-\lambda) \times S_a \qquad (2)$$

where S_v and S_a are express in term of the logarithm of the output probabilities. The final solution corresponds to the maximum of this bimodal output probability estimate.

We calculate the visual weighted factor λ as in (Adjouani and Benoît, 1996) with the following formula:

$$\lambda = \frac{\sigma_v}{\sigma_v + \sigma_a} \qquad (3)$$

In which σ_v and σ_a are respectively the dispersion of video and acoustic scores with the signification that low values of dispersion indicates high ambiguity in the decision coming from the corresponding component.

The value of the dispersion is calculated as follows:

$$\sigma_m = \frac{\sum_{\{i,j\} \subset \{1,2,3,4\}} \left| S_m^i - S_m^j \right|}{C_4^2} \qquad (4)$$

where S_m^i represent the logarithm of output probability corresponding to the i-th recognition hypothesis for the modality $m \subset \{a, v\}$.

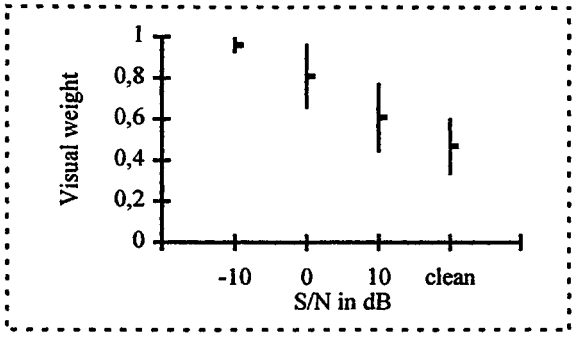

Figure 2: Variation of the visual weighting factor λ as a function of noise level in the SI based system

Figure 2 shows the variation of λ as a function of noise level. The corresponding mean and standard deviation are calculated over the test set.

The variation of the visual weight confirm the general hypothesis about the additive information included in lip shapes, especially in noisy environments. For example, the mean of the visual weight is about 0.5 in clean conditions and about 0.9 at -10 dB.

However, the estimation of the visual dispersion σ_v seem to be biased, since its value is calculated from the acoustic and visual scores corresponding to the N-best acoustic recognition hypotheses. This explain, we believe, the overestimation of the visual weighting factor λ.

3.3 Adaptive Weights for Hybrid Identification Model

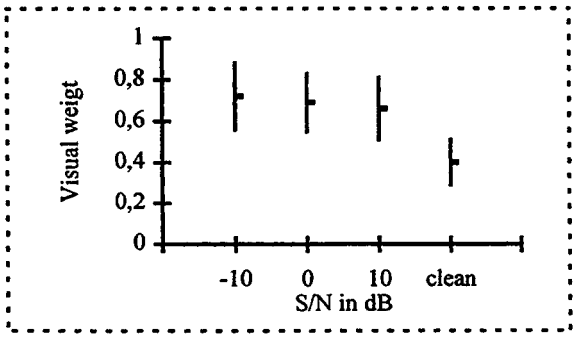

Figure 3: Variation of the visual weighting factor λ as a function of noise level in the DI+SI based system

The visual weight γ_V corresponding to the DI-based component has a fixed low value of 0.2, while the one λ corresponding to the SI-based component is calculated as previously.

We plotted in Figure 3 the mean and the standard deviation of λ. In non-noisy environment corresponding to high values of S/N, the mean of λ is about 0.4, which means that both modalities are taken into account with scarcely equal weights. In very noisy environment at -10 dB, the mean of λ is about 0.7, i. e. the visual source is preponderant for recognition process. The visual weights are less than in those corresponding to the SI based system, over different test condition. This may be explain by the fact that the acoustic-only component is replaced by a DI-based AV component.

4. EXPERIMENTS

4.1 Test Task and Results

System \ SNR	clean	10 dB	0 dB	-10 dB
Acoustic system	90.8 %	85.5 %	67.9 %	-44.3 %
DI based AV system				
equal weights	87.8 %	86.2 %	62.3 %	37.6 %
adaptive weights	95.4 %	88.3 %	75.0 %	37.6 %
SI based AV system				
equal weights	90.1 %	88.0 %	79.9 %	-32.3 %
adaptive weights	91.5 %	87.3 %	75.3 %	-32.7 %
DI+SI based AV system				
equal weights	93.6 %	88.0 %	76.7 %	40.4 %
adaptive weights	94.3 %	89.4%	78.1 %	40.4 %

Table 1: System performances with equal weights and automatic adaptive weights

The four systems were experimented on the same task: recognition of connected letters in French. The corpus, realised at ICP-Grenoble, it is composed of 200 utterances, of which two thirds were used for learning and one third for test. The acoustic signal is artificially degraded with dining-hall noise at a SNR of 10, 0 and -10 dB.

System performances are expressed in letter accuracy (correct letters minus inserted letters). The results obtained with the AV systems using equal weights and automatic adaptive modality weights, over different test condition, are shown in Table 1 and plotted in Figure 4.

4.2 Discussion

These results confirm the general hypothesis concerning the importance of weighting modality factor in the AV recognition process. Indeed, generally, system performances obtained with adaptive weights are better than those obtained with equal weights. However in the case of SI based AV system using adaptive weights instead of equal weights decrease its performance, except for clean data.

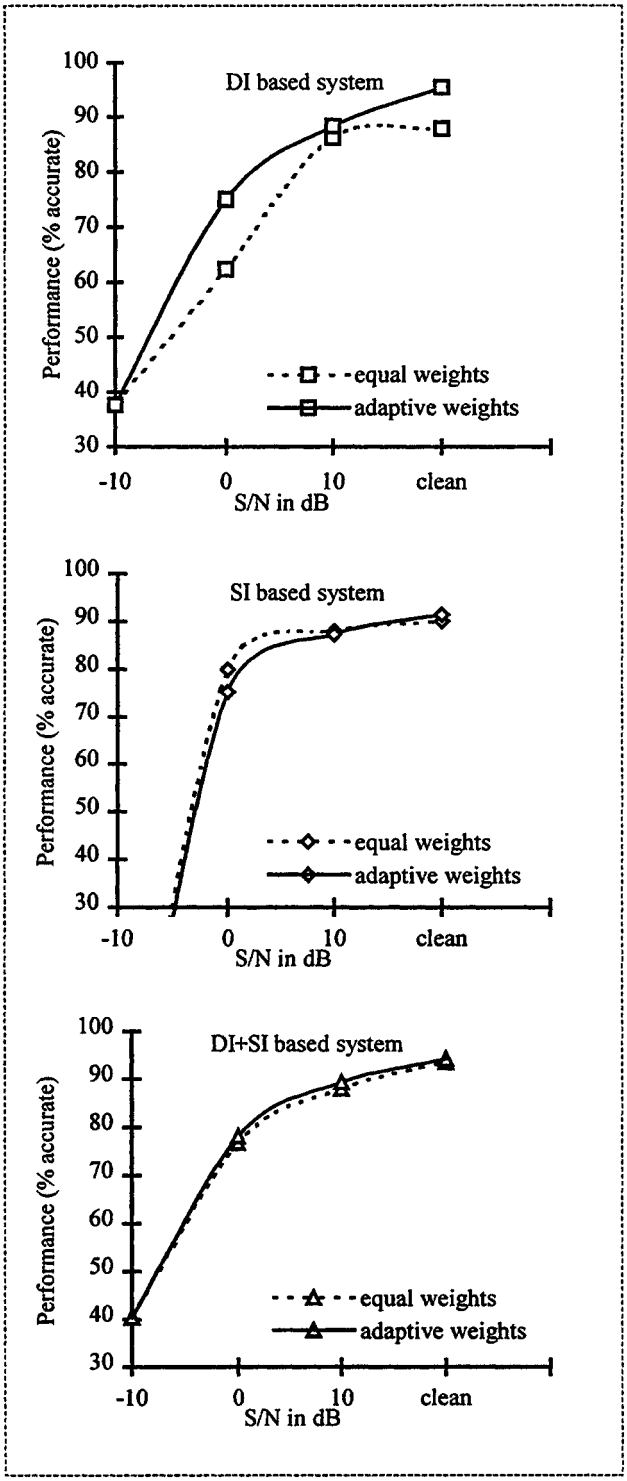

Figure 4: System performances with equal weights and automatic adaptive weights

We explain these results by the fact that the estimation of the visual dispersion σ_v is biased, which implies an overestimation of the visual weight λ.

Even if system performances are not as good as those corresponding to the empirically-obtained optimal modality weights (Alissali *et al.*, 1996) in some cases (clean and -10dB), they are generally better than those obtained with any fixed weights, over different test conditions. This confirm the appropriateness of automatic determination of audio and visual weights for AV recognition process.

5. CONCLUSION

In this paper we presented our work on adaptive integration of visual information for different AV recognition systems, which implement the DI model, the SI model or the hybrid DI+SI one. The acoustic and visual weights are dynamically adjusted manly according to the noise level.

The results we obtained are satisfactory especially for moderate noise level (10 dB and 0 dB) and confirm that the adaptive scheme proposed by (Adjouani and Benoît, 1996) is also well suited for more complex connected-letter-recognition task. However these results are yet to be confirm on more important corpus.

Further work is also to be done in order to exploit the auditory and visual confusion of phonemes for determining acoustic and visual weights.

6. REFERENCES

Adjouani A. and Benoît C. (1996), « On the Integration of Auditory and Visual Parameters in an HMM-based ASR », in *Speechreading by Humans and Machines*, Stork D. and Hennecke M., Eds., NATO ASI Series, Vol. 150, pp. 461-473.

Alissali M., Deléglise P. and Rogozan A. (1996), « Asynchronous Integration of Visual Information in an automatic Speech Recognition System », in *Proceedings of International Conference on Spoken Language Processing*, pp. 34-37, Philadelphia, USA.

Lallouache M. T. (1991), « Un poste visage-parole couleur », *Thèse de doctorat*, INPG-Grenoble, France.

Meier U., Wolfgang H. and Duchnowski P. (1996), « Adaptive Bimodal Sensor Fusion for Automatic Speechreading », in *Proceedings of International Conference on Acoustics, Speech and Signal Processing.*

Robert-Ribes J., Piquemal M., Schwartz J.-L. and Escudier P. (1996), « Exploiting Sensor Fusion Architectures and Stimuli Complementarity in AV Speech Recognition », in *Speechreading by Humans and Machines*, Stork D. and Hennecke M., Eds., NATO ASI Series, Vol. 150, pp. 193-211.

Silsbee P. and Su Q. (1996), « Audio-visual Sensory Integration using Hidden Markov Models », in *Speechreading by Humans and Machines*, Stork D. and Hennecke M., Eds., NATO ASI Series, Vol. 150, pp. 489-497.